Steps towards a GENIA dependency treebank

Schneider, G ; Rinaldi, Fabio ; Kaljurand, K ; Hess, M

Abstract: In this paper we describe on-going work aimed at creating a dependency-based annotated treebank for the BioMedical domain. Our starting point is the GENIA corpus, which is a corpus of 2000 MEDLINE abstracts, which has been manually annotated for various biological entities, according to the GENIA Ontology. There is an exponential growth of published research in this sector, which makes it difficult even for the experts to follow the recent developments. This creates the need for tools that can automatically process the research literature and extract only relevant information, such as interactions between genes and proteins. In order for these tools to be developed, annotated resources, such as corpora and Treebanks are of fundamental importance. Such resources will support the development of practical domain-specific information extraction tools.

Steps towards a GENIA Dependency Treebank

Gerold Schneider, Fabio Rinaldi, Kaarel Kaljurand and Michael Hess
Institute of Computational Linguistics, University of Zurich
{gschneid,rinaldi,kalju,hess}@ifi.unizh.ch

1 Introduction

In this paper we describe on-going work aimed at creating a dependency-based annotated treebank for the BioMedical domain. Our starting point is the GENIA corpus [14], which is a corpus of 2000 MEDLINE abstracts, which has been manually annotated for various biological entities, according to the GENIA Ontology.[1]

There is an exponential growth of published research in this sector, which makes it difficult even for the experts to follow the recent developments. This creates the need for tools that can automatically process the research literature and extract only relevant information, such as interactions between genes and proteins. In order for these tools to be developed, annotated resources, such as corpora and Treebanks are of fundamental importance. Such resources will support the development of practical domain-specific information extraction tools.

For an information extraction application extracting relations between genes and proteins [19] the dependency based parser Pro3Gres [20, 21] has been used. Pro3Gres is an open, modular and highly parameterized system. The module interaction can be seen in fig. 1. Pro3Gres is fast and robust, it parses the entire GENIA in under 3 hours. Although its performance is competitive, a considerable effort will have to go into correcting it to achieve a nearly error-free treebank.[2]

The creators of GENIA are currently planning to release a version of GENIA enriched with syntactic annotations based on a HPSG analysis of the corpus [22]. Our work can be considered parallel and complementary to theirs. We intend to compare and coordinate our results with the HPSG parsing based GENIA Treebank that is becoming available from the GENIA project. We also plan to make a dependency analysis widely available for research activities.

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/genia-ontology.html

[2] Whether this task can be be completed will depend on future funding.

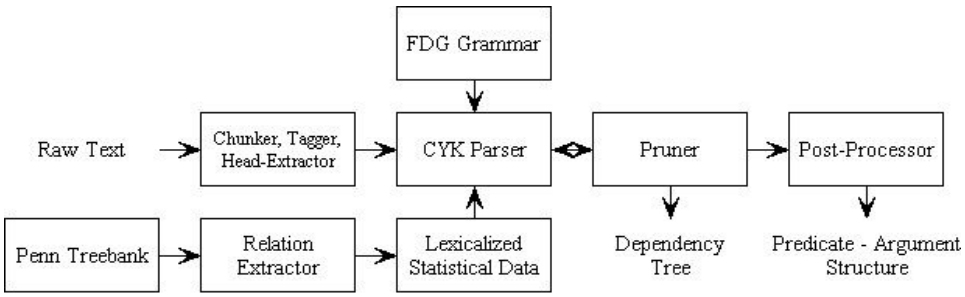

Figure 1: Pro3Gres flowchart

LTPos-tagged Parsing	Subject	Object	noun-PP	verb-PP
Precision	91.5	90.3	70.5	72.5
Recall	80.6	83.4	64.0	86.4

Table 1: Percentage results of evaluating the LTPos tagger based parser output on Carroll's test corpus on subject, object and PP-attachment relations

In this paper we show that the annotation effort can be reduced by using a high-performance parser. For this purpose, we test settings that allow us to optimize on recall and on precision respectively. The errors of the Pro3Gres parser are analyzed in detail. After an evaluation on a general corpus and on the GENIA corpus, we describe methods to minimize the annotator's task: high recall and high precision parsing. A practical evaluation discusses the impact of errors for obtaining domain knowledge and we conduct an analysis of remaining errors.

2 General Evaluation

[15] suggests evaluating on the linguistically meaningful level of syntactic relations. For the first evaluation, a hand-compiled gold standard following this suggestion is used [5]. It contains the grammatical relations of 500 random sentences from the Susanne corpus. Results are in table 1. The mapping between our and Carroll's annotation is discussed in [20].

A detailed analysis breaks down the errors into classes, in table 3 for PP-attachment, and in table 4 for subject and object precision. The analysis identifies mistagging and mischunking as generally important error sources in addition to parsing, but also that differing grammar assumptions are a problem. For the evaluation on the Carroll corpus, a mapping to our relation types was necessary [20]. Mapping one annotation scheme to another is non-trivial and can only lead to indicative results [11]. An obvious response to the big influence of tagging and chunking errors was to try a different tagger. Instead of LTPos [16] a Maximum

MaxEnt-tagged Parsing	Subject	Object	noun-PP	verb-PP
Precision	92.4	89.5	72.9	70.3
Recall	81.1	83.9	64.5	84.4

Table 2: Percentage results of evaluating Charniak's MaxEnt tagger based parser output on Carroll's test corpus on subject, object and PP-attachment relations

	Attachment Error	Head Extraction Error	Chunking or Tagging	compl/prep Error	Grammar Mistake or incompl. Parse	Grammar Assumption
Noun-PP Prec.	22	1	8	0	3	3
Verb-PP Prec.	12	1	5	1	1	2
Noun-PP Recall	25	1	14	0	12	5
Verb-PP Arg. Recall	2	0	1	0	0	0
Total	61 51%	3 3%	28 24%	1 1%	16 13%	10 12%

Table 3: Error Classification of PP-Attachment errors from the first 100 Carroll corpus sentences

Entropy tagger has been tested [6]. The results are in table 2. The comparison yields no clear result, but has prompted us to do the first two high-precision experiments in section 4.

For the second evaluation, 100 random GENIA sentences have been manually annotated and compared to the output of the parser[3]. We keep the manual annotation for multi-word biological terms as chunked input to the parser[4]. The results are in table 5.

3 High Recall Parsing

The annotation task is greatly facilitated if the annotator, instead of being asked to annotate every sentence manually, can choose from a (relatively short) ranked list of analyses. [3] have shown that parser-assisted annotation (in their case an

[3]This is a small set. Average sentence length is 17.9 chunks, compared to 17.0 in the whole GENIA, so we can assume that it is fairly representative

[4]see [19] for the impact of keeping this information

	Spurious Error	Chunking or Tagging	Control	Parsing Error	Rel. Pronoun Resolution	Grammar Mistake or incompl. Parse	Grammar Assumption
Subject Precision	8	22	9	15	4	9	9
Object Precision	0	12	1	5	1	1	2
Total	8 8%	34 35%	10 10%	20 20%	5 5%	10 10%	11 11%

Table 4: Error Classification of Subject and Object Precision errors of all Carroll corpus sentences

Percentages on GENIA	Subject	Object	noun-PP	verb-PP	subord. clause
Precision	90	93	85	82	68
Recall	87	91	82	84	73

Table 5: Evaluation of 100 sentences of the GENIA corpus, using multi-word term boundary information

	Carroll				GENIA			
High Recall	Subject	Object	noun-PP	verb-PP	Subject	Object	noun-PP	verb-PP
1 analysis	80.8	83.4	64.9	86.4	86.6	91.1	81.6	83.3
max. 2 analyses	81.4	83.6	70.4	89.9	87.7	91.1	85.4	83.3
max. 4 analyses	81.6	84.1	73.9	90.4	90.3	91.1	91.8	86.2
max. 8 analyses	81.8	84.1	75.2	91.4	91.3	91.1	93.7	86.2
max. 16 analyses	81.9	84.4	75.4	91.4	91.8	91.1	94.2	86.2

Table 6: Percentage results of recall among first N-ranked analyses

interactive scenario with a shallow parser [2]) greatly increases annotation speed. Table 6 shows the increase in recall in relation to the length of the list of analyses. Lists longer than 16 readings of a sentence (which convey 4 2-way ambiguous relations) are thought to be prohibitively long for manual scanning.

The $subj$, obj and the two PP-relations together average above 90% recall in GENIA, which means that less than one in ten of these relations need to be added manually by the annotator. Generally, recall in GENIA is higher. This is due to the following reasons:

1. As we have annotated our test corpus with the Pro3Gres scheme, there are no spurious mapping errors
2. We can profit from the fact that GENIA contains near-perfect tagging and multi-word term (MWT) information
3. We have written an unsupervised learning module and applied it to GENIA. Based on the fact that sentence-initial <NP PP*> sequences are almost always unambiguous [8], it learns which nouns are allowed to be modified by several PPs and restricts noun modification by several PPs accordingly. This especially explains the very high noun-PP-attachment recall.

4 High Precision Parsing

In order to keep the necessity for intervention of a human annotator during corpus annotation to a minimum, it is desirable to recognize a maximum number of unproblematic relations. An alternative annotation scenario is thus to report the highest ranked parse and to point out to the human annotator the few difficult and

Experiment 1	Subject	Object	noun-PP	verb-PP
Precision	92.2	95.4	85.6	71.6
Recall	31.5	30.7	23.2	27.8

Table 7: Percentage results of Experiment 1: keeping only sentences with identical tags from two taggers, on Carroll's test corpus on subject, object and PP-attachment relations

Experiment 2	Subject	Object	noun-PP	verb-PP
Precision	94.1	93.0	73.3	75.4
Recall	76.4	78.8	60.5	80.3

Table 8: Percentage results of Experiment2: keeping only agreeing relations arising from parsing with two taggers, on Carroll's test corpus on subject, object and PP-attachment relations

highly ambiguous relations in a given analysis. Parsing methods that optimize precision while reducing recall up to an acceptable point are required. A related study on this subject is [4]. This field of research may eventually lead to the automatized detection of potential parsing errors. It is also important for building up knowledge databases automatically, where recall deficiencies are often compensated by natural language redundancy, but asserting wrong knowledge arising from low precision poses a serious problem.

Experiment 1: Tagger Agreement Different taggers often make different mistakes. In a simple experiment, only sentences where both taggers deliver identical tags are used. Precision increases, but the large cost of decrease in recall is unacceptable, as shown in table 7.

Experiment 2: Grammatical Relations Agreement when using different Taggers In order to minimize the loss in recall in the previous experiment, the output of each tagger is used as input to the LTChunk chunker and the Pro3Gres parser. Only grammatical relations that are different due to the tagging differences are discarded. The increase in precision is similar to experiment 1 (noun PP-attachment is slightly worse) while the decrease in recall is much more moderate, as table 8 shows.

Experiment 3: Parsing Alternatives Agreement In this experiment, the relation intersection between the 2 top ranked analyses is kept. This amounts to discarding only the most ambiguous relation of any given sentence. The decrease in recall (table 9) is higher than in experiment 2. Mainly the PP-attachment relations profit, which are often the most ambiguous relations, and which are more affected

Experiment 3		Subject	Object	noun-PP	verb-PP	subord. S
Carroll	Precision	92.6	90.1	76.6	76.7	68.2
	Recall	76.8	63.6	53.7	67.2	n/a
GENIA	Precision	91.1	93.4	87.0	84.2	65.2
	Recall	78.1	65.8	68.0	70.5	60.4

Table 9: Percentage results of Experiment 3: discarding the most ambiguous relation in each sentence, for subject, object, PP-attachment and subordinate sentence relations

by attachment ambiguities than other relations.

Experiment 4: Trust Short Distances Relation spanning short distances are intuitively thought to be easier for the parser to find. Experiment 4 discards all relations that are longer than a certain threshold. Length is measured in chunks. The experiment has been conducted at several distances for the Carroll test corpus (table 10) and for the 100 manually annotated GENIA sentences (table 11).

The results reveal interesting differences between different relation types. For $subj$, longer distances are almost as reliable. obj relations are almost exclusively very short. Subordinate clause relations are difficult and mostly very long, about 20% spanning at least 5 chunks. For envisaged applications, e.g. protein interaction relations, sentence subordination is less important. PP-attachment relations very strongly depend on distance. This is largely due to the fact that many PP-attachments across longer distances[5] are in competition with intervening other PPs and thus exponentially lower the baseline[6].

When comparing the two evaluation corpora and genres a major difference is PP-attachments. The complexity of medical language partly stems from very complex nouns with embedded PPs (see e.g. fig. 2). The noun-PP-attachment per sentence ratio is 2.1 in our GENIA 100 test corpus and 1.6 in Carroll. The fact that the performance on GENIA is better than on Carroll can largely be explained by our remarks in section 3.

Experiment 5: Cut low probability parsing decisions In a first attempt, experiments with an increased probability cutoff at parse time were conducted. However, they had the effect of greatly increasing the amount of non-full parses, thus returning many local analyses that the syntactic parsing context would have disambiguated. Precision remained comparable, while recall dropped. In a second

[5]observe that "longer distances" does not entail a long-distance dependency traditionally expressed by coindexing or movement, although a considerable portion of the "longer distances" here are long-distance dependencies, for example fronted PPs attaching to the verb

[6][7] describe for PP attachment that a sequence <verb-NP-PP*> with n PPs has C_{n+1} analyses, where C_{n+1} is the $(n+1)$'th Catalan number. The Catalan number C_n is defined as $\frac{1}{n+1}\binom{2n}{n}$

```
Interaction_NN of_IN nuclear_JJ extracts_NNS from_IN various_JJ cell_NN lines_NNS and_CC tissue_NN
with_IN the_DT MNP_NN site_NN leads_VBZ to_TO the_DT formation_NN of_IN fast-migrating_JJ
protein-DNA_JJ complexes_NNS with_IN similar_JJ but_CC distinct_JJ electrophoretic_JJ mobilities_NNS

prep('extract#3', 'of#2', _, '(<-)').              prep('line#5', 'from#4', _, '(<-)').
conj('tissue#7', 'and#6', _, '(<-)').              conj('line#5', 'tissue#7', 'and#6', '(->)').
prep('site#9', 'with#8', _, '(<-)').               modpp('line#5', 'site#9', 'with#8', '(->)').
modpp('extract#3', 'line#5', 'from#4', '(->)').    modpp('interaction#1', 'extract#3', 'of#2', '(->)').
subj('lead#10', 'interaction#1', _, '(<-)').       prep('formation#12', 'to#11', _, '(<-)').
prep('complex#14', 'of#13', _, '(<-)').            modpp('formation#12', 'complex#14', 'of#13', '(->)').
pobj('lead#10', 'formation#12', 'to#11', '(->)').  prep('mobility#16', 'with#15', _, '(<-)').
pobj('lead#10', 'mobility#16', 'with#15', '(->)').
```

Figure 2: A sample sentence with its top-ranked grammatical relation annotation

Experiment 4 on Carroll		Subject	Object	noun-PP	verb-PP	subord. S
Distance 1-2	Precision	94.3	90.5	76.0	85.7	74.1
	Recall	70.5	83.9	52.3	69.7	n/a
Distance 1-3	Precision	92.7	90.3	74.0	77.5	74.7
	Recall	75.5	84.1	59.2	78.3	n/a
Distance 1-4	Precision	92.2	90.0	73.5	75.2	70.8
	Recall	76.8	84.4	61.7	81.3	n/a
Distance 1-5	Precision	92.3	89.8	73.3	74.2	69.1
	Recall	78.6	84.4	62.5	82.3	n/a
Distance > 5	Precision	96.0	null	0.0	37.4	55.0
	Recall	5.4	null	0.0	2.0	n/a

Table 10: Percentage results of Experiment 4: discarding relations that span long distances, on Carroll's test corpus

approach, the parsing algorithm remains unchanged, but only relations whose probability is above a certain threshold are reported. Here we profit from the fact that the Pro3Gres probabilities express decision probabilities at each given ambiguous point as suggested by [10]. In addition to offering a psycholinguistically plausible model this has the advantage that points of uncertain decisions and high entropy can be directly pinpointed. These experiments have been made on the highly ambiguous PP-attachment relations, see table 12.

Below threshold values of about 0.5 there is a reasonable trade-off in gained precision for lost recall. With higher thresholds, precision stagnates while recall drops off.

Combinations Most of the above high-precision experiments can be combined in various ways. E.g. combinations of experiment 3, 4 and 5 are reported in tables 13 with threshold 0.4 and distances 1 to 5. This sample combination on the GENIA annotation task allows us to reach about 9 out of 10 precision at 2 out of 3 recall for all reported relations.

5 Practical Evaluation

Our interest lies in the discovery of domain specific relations, such as "Protein *activates* Gene". Most of the NLP techniques applied to the domain of molecular

Experiment 4 on GENIA		Subject	Object	noun-PP	verb-PP	subord. S
Distance 1-2	Precision	92.3	92.9	88.1	95.5	75.0
	Recall	57.1	91.1	79.0	64.7	14.5
Distance 1-3	Precision	89.5	92.9	87.2	87.6	84.0
	Recall	64.8	91.1	79.0	74.1	39.6
Distance 1-4	Precision	90.2	92.9	86.8	87.5	79.3
	Recall	69.4	91.1	79.5	77.7	43.8
Distance 1-5	Precision	90.9	92.9	85.6	85.6	71.8
	Recall	74.5	91.1	80.0	79.1	54.2
Distance > 5	Precision	89.3	null	0.0	41.7	57.1
	Recall	2.0	null	0.0	3.6	18.7

Table 11: Percentage results of Experiment 4: discarding relations that span long distances, on GENIA corpus relations

Experiment 5		Carroll		GENIA	
		noun-PP	verb-PP	noun-PP	verb-PP
Threshold 0.3	Precision	73.7	71.0	84.5	81.6
	Recall	64.2	84.8	79.0	82.7
Threshold 0.4	Precision	74.4	71.3	85.3	81.3
	Recall	63.6	84.3	78.1	80.5
Threshold 0.5	Precision	76.0	72.6	86.2	79.8
	Recall	61.3	81.3	72.4	71.2
Threshold 0.6	Precision	76.6	72.8	87.4	82.3
	Recall	56.2	73.2	70.4	59.7
Threshold 0.7	Precision	77.0	72.5	87.6	81.5
	Recall	52.6	66.1	68.6	51.8
Threshold 0.8	Precision	77.0	73.0	88.1	80.3
	Recall	51.2	63.1	64.8	45.3
Threshold 0.9	Precision	77.1	73.6	88.2	79.7
	Recall	50.9	62.1	64.8	43.9

Table 12: Percentage results of Experiment 5: discarding low-probability relations, on Carroll's and the GENIA test corpus

biology focus on the discovery of Entities, such as Genes and Proteins, (see for instance [1]). However there are also interesting applications aiming at detecting syntactic and semantic relations among those entities. Examples of systems aiming at detecting relations are the following:

- [9] identifies possible drug-interaction relations between proteins and chemicals using a "bag of words" approach applied to the sentence level.
- [17] reports on extraction of protein-protein interactions based on a combination of syntactic patterns.
- [12] describes a system (GENIES) which extracts and structures information about cellular pathways from the biological literature.
- [18] processes titles and abstracts of Medline articles focusing on relation identification (in particular the `inhibit` relation)

Experiments	Carroll				GENIA			
3,4,5 combined	ubject	Object	noun-PP	verb-PP	Subject	Object	noun-PP	verb-PP
Precision	92.6	90.1	78.9	80.5	92.4	93.5	87.9	88.1
Recall	75.0	63.4	51.2	67.2	67.3	67.0	66.7	65.5

Table 13: Percentage results of Experiments 3, 4 and 5 combined at threshold 0.4 and distances 1 to 5

- [13] uses a template-based Information Extraction approach, focusing on the roles of specific amino acid residues in protein molecules

In order to discover domain specific relations we believe that an accurate detection of predicate/argument relations is essential. We have asked domain experts to evaluate the quality of the extracted relations, so far focusing on triples of the form (predicate - subject - object).[7]

A first evaluation was based on assigning a simple key code to each record: 'P' for positive (biologically relevant and correct, 53 cases), 'Y' for acceptable (biologically relevant but not completely correct, 102 cases) and 'N' (not biologically relevant or seriously wrong, 14 cases). This result was considered as encouraging as it showed 91.7% of relevant records.

On closer inspection of the expert results, we identified a number of 'typical cases', which we then asked the expert to evaluate in detail. In this second evaluation the expert had to evaluate each argument separately and mark it according to the following codes:

- [Y] the argument is correct and informative
- [N] the argument is completely wrong
- [Pr] the argument is correct, but it is a pronoun, and it would need to be resolved to be significant (e.g. "This protein").
- [A+] the argument is "too large" (which implies that a prepositional phrase has been erroneously attached to it)
- [A-] the argument is "too small" (which implies that an attachment has been omitted)

Despite parsing errors – some of which we are now correcting in the parser – the results can be considered satisfactory, as they show 86.4% and 58.6% correct results in the detection of subjects and objects (respectively). If all loose cases are considered as positive (excluding only the 'N' cases), these results increase to 93.5% and 99.4% (respectively).

[7]This evaluation has been performed in collaboration with Biovista (`http://www.biovista.com/`)

	Y	N	Pr	A+	A-
Subject	146	11	4	6	2
Object	99	1	4	59	6

Table 14: Distribution of GENIA parsing errors in the application-oriented evaluation

Recall Error Classification on GENIA High Recall Parsing									
	Adjective Trans.	Incompl. Grammar	Chunking Error	Tagging Error	Incompl. Parse	LDD Resol.	Annotation Problem	Attachment Error	Conjunction Error
Subject Recall	1	2	1	2	0	6	2	1	1
Object Recall	1	1	2	0	0	0	1	1	1
N-PP Recall	0	2	2	0	3	0	2	3	0
V-PP Recall	1	5	2	4	2	1	2	2	0
Total	3	10	7	6	5	7	7	7	2

Table 15: Analysis of recall errors on max. 16 GENIA high recall parsing

Let us consider a realistic annotation scenario using the high recall parsing method from section 3 with the annotator selecting the best of top 16 analyses. Over subject, object and PP-attachment relations, recall is 564/618 = 91.3%. 54 errors stemming from 34 sentences remain. Table 15 breaks down these errors into classes.

Bearing in mind that the annotation problem errors are spurious errors, that long-distance dependencies (LDDs) are often left underspecified by statistical parsers, and that the parser is affected by tagging and chunking mistakes, actual high recall parsing performance for the evaluated relations can be confirmed to reach 95%.

Pro3Gres is a modular system. Tagging and chunking are external processes whose output can be confirmed or corrected by the user. We are now investigating ways to integrate annotator feedback at an interactive and especially at the post-parsing stage. The latter triggers re-parsing erroneous sentences using the annotator's safe corrections.

6 Conclusion

We have shown that the annotation effort for building a syntactically analyzed corpus can be reduced by using a high-performance deep-linguistic parser and that it is possible to pin-point places of high entropy, to optimize on recall or on precision, respectively, to distinguish between more and less reliable relations.

We have shown that Pro3Gres can do full, deep-linguistic parsing of BioMedical texts at competitive speed and accuracy. The parser's errors have been analyzed in detail. We plan to compare and coordinate our grammatical relations output to the GENIA Treebank that is becoming available from the GENIA project.

References

[1] Sophia Ananiadou and Jun'ichi Tsujii, editors. *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, 2003.

[2] Thorsten Brants. Cascaded markov models. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, pages 118–125, Bergen, Norway, 1999. University of Bergen.

[3] Thorsten Brants and Oliver Plaehn. Interactive corpus annotation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece, 2000.

[4] John Carroll and Ted Briscoe. High precision extraction of grammatical relations. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, Taipei, Taiwan, 2002.

[5] John Carroll, Guido Minnen, and Ted Briscoe. Corpus annotation for parser evaluation. In *Proceedings of the EACL-99 Post-Conference Workshop on Linguistically Interpreted Corpora*, Bergen, Norway, 1999.

[6] Eugene Charniak. A maximum-entropy-inspired parser. In *Proceedings of the North American Chapter of the ACL*, pages 132–139, 2000.

[7] K. Church and R. Patil. Coping with synactic ambiguits or how to put the block in the box on the table. *American Journal of Computational Linguistics*, 8(3-4):139–149, 1982.

[8] Michael Collins and James Brooks. Prepositional attachment through a backed-off model. In *Proceedings of the Third Workshop on Very Large Corpora*, Cambridge, MA, 1995.

[9] M. Craven and J. Kumlien. Constructing biological knowledge bases by extracting information from text sources. *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB-99)*, 1999.

[10] Matthew Crocker and Thorsten Brants. Wide coverage probabilistic sentence processing. *Journal of Psycholinguistic Research*, 29(6):647–669, 2000.

[11] Richard Crouch, Ronald M. Kaplan, Tracy H. King, and Stefan Riezler. A comparison of evaluation metrics for broad-coverage stochastic parsers. In *Beyond PARSEVAL workshop at 3rd Int. Conference on Language Resources an Evaluation (LREC'02)*, Las Palmas, 2002.

[12] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(1):S74–S82, 2001.

[13] R. Gaizauskas, G. Demetriou, P. J. Artymiuk, and Willett P. Protein Structures and Information Extraction from Biological Texts: The PASTA System. *Bioinformatics*, 19:135–143, 2003.

[14] J.D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(1):180–182, 2003.

[15] Dekang Lin. A dependency-based method for evaluating broad-coverage parsers. In *Proceedings of IJCAI-95*, Montreal, 1995.

[16] Andrei Mikheev. Automatic rule induction for unknown word guessing. *Computational Linguistics*, 23(3):405–423, 1997.

[17] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(2):155–161, 2001.

[18] J. Pustejovsky, J. Castaño, J. Zhang, B. Cochran, and M. Kotecki. Robust relational parsing over biomedical literature: Extracting inhibit relations. In *Pacific Symposium on Biocomputing*, 2002.

[19] Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, James Dowdall, Christos Andronis, Andreas Persidis, and Ourania Konstanti. Mining relations in the genia corpus. In Tobias Scheffer, editor, *Accepted for publication in: Second European Workshop on Data Mining and Text Mining for Bioinformatics*. ECML/PKDD, September 2004.

[20] Gerold Schneider. Extracting and using trace-free Functional Dependencies from the Penn Treebank to reduce parsing complexity. In *Proceedings of Treebanks and Linguistic Theories (TLT) 2003*, Växjö, Sweden, 2003.

[21] Gerold Schneider, James Dowdall, and Fabio Rinaldi. A robust and deep-linguistic theory applied to large-scale parsing. In *Coling 2004 Workshop on Robust Methods in the Analysis of Natural Language Data (ROMAND 2004)*, Geneva, Switzerland, August 2004, 2004.

[22] Akame Yakushiji, Yuka Tateisi, Yusuke Myao, and Jun'ichi Tsujii. Building the GENIA Dependency Grammar Treebank of BioMedical documents. In *Proceedings of ACL04, poster session*, 2004.